跨媒体分析与检索技术研究进展

王树徽 黄庆明

摘要 在网络跨媒体应用迅速兴起，网络内容对网络用户影响日益深刻的背景下，本文介绍了跨媒体分析与检索的相关理论和方法，包括如何提取网络跨媒体数据的多源自然属性和社会属性，揭示海量跨媒体的语义多样性及数据关联和内在信息传播机制，内容涵盖以下几方面：首先，讨论网络跨媒体数据的跨平台、多模态和来源广泛等特性及其带来的挑战和机遇，介绍跨媒体分析技术的特点和传统单一媒体分析的不同之处，以及跨媒体可能带来的科学和社会影响力；接下来，分别从跨媒体语义分析与理解、跨媒体关联建模和跨媒体社群分析等三个方面介绍跨媒体分析与检索技术的国内外研究现状；最后，介绍中科院计算所智能信息处理重点实验室在跨媒体语义分析理解，跨媒体中热点事件和话题分析以及跨媒体用户行为分析等方面的研究情况。

关键词：网络跨媒体，语义分析与理解，热点事件和话题分析，跨媒体用户行为分析

1　介绍

随着互联网技术和多媒体技术的飞速发展，互联网正在越来越深入人们的生产、生活、娱乐和社会交往等活动当中。近年来，网络多媒体和移动多媒体用户的数量呈现飞速增长态势，社交网络等新兴媒体在网络用户群体的使用率也接近甚至超过 50%。同时，文本已不再是信息和知识的唯一载体，知识的传播正在以更为灵活、多样、丰富和翔实的方式进行，信息与知识呈现多来源化，跨媒介化以及关联多样化等种种特性。另一方面，随着交互式网络技术的飞速发展，微博、图像视频分享网站、社交网络等诸多平台的兴起与普及，越来越多的用户在网络上以发布消息、张贴图片视频等方式传播消息、表达观点，通过与其他用户的信息交互机制获取大量知识。

网络数据除了呈现海量性特点之外，数据之间的关联性也在不断增强。这种关联性也成为网络信息除了自身内容之外的另外一个重要来源。在文本搜索领域，互联网搜索引擎公司谷歌（Google）利用的 PageRank 技术，通过分析和利用网页内容之间的超链接信息对网页的重要性进行计算，为海量网络内容检索带来了革命性的突破。与文本相比，网络多媒体数据之间的关联性较之一般的文本网页更加丰富。例如，网络图像和视频一般与大量的环绕文字共同出现，这些环绕文字提供了对视觉内容的描述性信息。由于交互式网络技术的兴旺发展，网络用户可对跨媒体进行编辑和标注，对视觉内容提供标注信息，可以自由转载、分享和评论跨媒体内容。如何有效地分析利用这类信息，成为多媒体信息检索领域研究的核心问题。

总体而言，网络信息越来越呈现海量、来源广泛、跨媒介、复杂关联等特性，数据与用户之间存在密不可分的互动关系。这些来自不同平台的不同类型的媒体和与之相关的社会属性信息更加紧密地混合在一起，以一种崭新的形式，更为形象地表示综合性知识，反映个体或者群体的社会行为。这种新的称为“跨媒体(Cross-media)”的媒体表现形式呈现出如下三个基本属性：

- **固有的跨模态和跨平台属性** 即文字、图像、视频、声音和链接等结构化或非结构化的跨模态属性及不同网络平台的数据之间的物理连接和高度相关的内容的多态

多义性。

- **丰富的表达和呈现力** 通过各类网络平台，跨媒体数据所固有的多源信息从不同角度来展示客观世界及其包含知识，而且多种表达形式形成的互补协同的描述具有更丰富的呈现能力。
- **媒体数据的社会性** 来源于不同渠道的各类跨媒体数据通过不同方式被赋予地理、时空、社区、热度、偏好等属性，数据与数据以及数据与用户之间交叉关联，通过群体推动的机制传播并动态演化，从而对媒体内容的语义理解产生主观影响。

跨媒体数据的模态包括文本信息、视觉信息、听觉信息，跨媒体数据的来源包括网页、网络视频、网络图像、社会媒体中的分享、转载、评注、引用、用户 GPS 轨迹信息、地点定位信息等。在网络的使用者对互联网的信息获取的依赖性不断加强的同时，他们自身也产生了大量的网络数据。由于网络数据的爆炸式增长，如果没有强有力的内容分析工具的帮助，用户很难从海量数据中获得所需的有用信息和知识。在本文中，我们将跨媒体中蕴含的知识分为两种主要类型：第一种是跨媒体的自然属性，即描述跨媒体的产生时间(When)、地点(Where)、描述什么内容(What)、如何发生(How)等方面的特性；第二种是跨媒体的社会属性，即与哪些人相关、影响作用了哪些人以及被哪些人的行为影响和作用。跨媒体分析和检索研究的目的之一，就是为了有效地提取这些属性，从而更好地认识跨媒体的产生、发展、传播和演化机制。

有别于传统的结构化和非结构化数据，跨媒体数据往往融合了多种模态信息。跨媒体的多源属性也导致了信息的交叉传播与整合，体现了多个平台、不同媒体和用户群体间的合作、共生、互动与协调。对于同一个跨媒体热点事件或者话题，新闻网站里往往对事件的发生以及后续影响发展进行深入报道，而社交媒体更多体现了民众对事件的关注度以及主观反响。例如，针对近期发生的利比亚政变，世界各大主流媒体和电视网对其进行深入的全程跟踪报道。在网络上，瞬时出现很多针对此次事件的纪实资料和评论分析，大多数采用文字、图片和视频等图文并茂的方式呈现。在社会媒体中，普通用户对信息进行引用、转载和评述，一时间在网络上针对这次事件讨论的热度不断升高，并一度呈现白热化态势。在这次政治事件中，传统媒体和新兴媒体参与信息传播的方式和手段，以及事件发展的信息传播和演化过程，再一次让我们对跨媒体信息传播有了直观形象的认识。

跨媒体带来丰富信息量的同时，也为媒体分析与理解的相关研究带来了新的挑战。传统的针对单一类型、小数据量数据的分析方法已经不能满足技术需求。针对多源异构大数据计算的研究已受到各国的充分重视。在我国，多源异构大数据和网络跨媒体数据分析被列入国家重点基础研究发展计划和重大科学研究计划 2014 年重要支持方向，对满足国家的重大战略以及商业应用需求具有重要作用。从目前的技术发展情况看，跨媒体分析领域的研究主要存在如下趋势和特点：

首先，对海量媒体内容进行有效理解，是实现众多跨媒体应用的先决条件。由于海量跨媒体数据的高维、多源和异构特性，存在极为丰富的信息量的同时，也不可避免地存在巨大的语义鸿沟，从而需要对符合用户高层认知的信息进行分析和提取。例如对跨媒体内容进行分类、标注、聚类等。然而，针对跨媒体理解问题，以往的研究大多只考虑单特征或者单一模态，例如纯文本或者纯视觉信息，且所关注的语义集合十分有限，所采用的模型一般也为浅层分析模型，不能很好地弥合底层特征与高层语义之间的巨大鸿沟，不利于处理开放网络环境的海量跨媒体的语义学习问题。为应对这些挑战，近年来在相关研究领域提出的多特征融合、特征学习和相关学习等新思路和新方法，为跨媒体分析与检索提供了新的研究思路，

有助于解决跨媒体数据的复杂分布和语义鸿沟问题。

第二，跨媒体检索是新一代媒体内容服务的趋势之一。对跨媒体检索技术的迫切需求，可以从两个方面来概括。首先，由于跨媒体数据的不断涌现，网络用户早已不满足于检索和浏览单一形态的媒体内容，而往往希望通过更加灵活的方式对信息进行查找和搜集。例如，用户希望通过输入一些文本查询，找到具有相关内容的网页、视频、图像和音频等，或者通过输入一幅素描的长城，检索关于长城的自然图像或者油画等。如何根据用户的任意输入查询来查找及定制不同来源的多种模态的媒体信息，已经成为迫在眉睫的问题。另一方面，未来以人为中心的数据检索技术应能对任意类型的输入进行处理，并准确理解用户意图，正确返回用户感兴趣的目标跨媒体数据。为达到上述目的，其关键在于如何建立不同模态、不同来源数据的具有语义一致性的可度量紧凑表示。通过融合跨媒体数据的多源信息（例如：内容共生性信息、语义标注信息、超链接信息、社会信息等），构建跨媒体数据的多源知识表示模型，构建有利于有效学习的跨媒体语义一致性度量表示。

第三，由于跨媒体数据的海量性和用户偏好的多样性，媒体信息的个性化定制是信息内容交换、共享和管理的核心问题之一。随着以社交媒体为代表的网络信息分享网站的崛起和涌现，每时每刻都会有数以万计的各种媒体信息在网络上出现和传播。普通民众从信息的接收者变成了数据和网络话题的制造者和直接参与者，并通过各类网络应用连结在一起形成网络群体连接关系。这种关系包含现实生活在网络上的延伸，也包含因为拥有相同而明确的目标和期望而关联起来的纯虚拟群体。社群的形成往往建立在共同的兴趣、喜好背景或者对某种事物的共同认知或关注上，因而社群内的成员往往具有某些相似或关联属性，例如对跨媒体内容的认知喜好、对网络事件的观点看法等。如何根据对用户的属性、行为和意图分析，从海量的数据中找到所需要的目标内容，是一个非常具有挑战性的难题。

综上所述，跨媒体的兴起，为新一代网络多媒体检索提供了前所未有的发展机遇。以往专注于多媒体自身内容分析的研究思路已不能很好地适应跨媒体数据的跨模态、跨平台等多源属性，不能有效利用数据之间的关联关系，对跨媒体内容进行更为深入的内容理解和更准确的检索。另一方面，由于跨媒体数据所固有的社会属性反映了跨媒体数据本身与网络社群用户之间的紧密关联关系，这为研究更加人性化和个性化的跨媒体检索技术提供了很好的契机。针对跨媒体的数据多源性、跨模态性、海量性及分布复杂且不均衡等特点，研究有效的跨媒体语义分析和检索技术，对网络社群行为进行建模分析，充分挖掘跨媒体信息处理和网络社群用户行为之间的关系，可为海量跨媒体信息处理提供新的解决方案。从应用角度讲，这又会为个性化检索、推荐、内容定制提供契机，为更有效地进行内容推送、广告投放、资讯发布给予指导。从社会角度来说，跨媒体分析为网络的内容过滤和网络社群行为分析提供强有力的支持，有助于维护社会公共安全，促进社会公平正义，保持社会良好秩序。

2 国内外研究现状

由于网络和多媒体技术的不断发展，网络多媒体数据呈现爆炸性增长趋势。对多源跨媒体数据智能处理已经受到国内外学者的广泛关注，近年来涌现了大量的研究成果。跨媒体分析涉及的领域较多，例如：多媒体分析、计算机视觉、自然语言处理、音频分析、网页分析、社会网络分析等等。本文将从三个跨媒体的核心分析对象（语义、关联、社群）来对相关工作进行剖析。

2.1 海量跨媒体数据的语义分析与理解

跨媒体数据体量巨大，内容丰富多样。其中（尤其是视觉数据）蕴含的语义信息对于跨

媒体分析理解起到至关重要的作用。对于视觉数据，特征表示往往直接影响模型的最终性能。然而，受制于图像底层特征和高层语义之间的语义鸿沟[1]，图像的类别信息很难直接从视觉底层特征直接获得。另一方面，现有的不同视觉底层特征一般从具体的某一方面（例如颜色、纹理和形状信息）描述视觉内容[2,3]。不同的底层特征对不同类别的图像识别的贡献不尽相同。即使对于某个典型主题的图像内容，不同的表现形式以及白天、黑夜等不同光照条件，在带来不同的感官感受的同时也由于其所具有的丰富视觉内容造成了网络图像检索、分类模型学习的困难。研究者致力于通过设计特征的提取来解决上述问题。虽然在一些情况下这些特征显示了充分的效果，但是在大多数情况下仍然存在判别力不足的问题，并不能用来解决识别、检测等涉及相对高层语义的问题。近年来，学者提出一种基于稀疏编码的局部视觉单词编码方法[36]，在多个基准视觉数据集上获得优越的分类性能。基于稀疏编码的思想，学者们还提出了若干类似的方法，例如局部线性编码[37]等，也都被证明了能比传统的视觉特征更好地应对视觉表观信息丰富的变化。马瑞艾尔（Marial）等人[38]进一步发现将判别信息（例如类别信息等）引入稀疏编码过程，能够使所提特征具有更好的语义一致性。这类方法也为相关研究提供了指导性信息。

对于不同语义主题的图像，由于内容既存在类内的变化，也存在一定的类间差异及共性，类别间的组织结构对分类识别模型的学习起到重要的作用。传统的一对多的分类模型虽然成功应用于处理小数据量或者理想实验环境数据[34-36]，但由于极多类别带来的类样本分布极度不均衡以及数据来源域的多样性，造成了模型的退化。一种可行的解决之道是利用图像类别的层次化组织关系[40,43]构建判别模型。近年来，深度学习[41,42]被广泛应用在视频、图像、音频、文本等数据分类和处理，并获得了超越（几乎所有）经典方法的性能，已逐渐成为一种基准方法。深度学习对数据进行多个层次的“抽象”表示，这与以往统计学习方法具有显著不同，更适合于处理具有复杂内容的跨媒体数据，将成为研究的热点。

作为另外一种可行途径，利用多个核函数处理多特征信息的多特征融合方法在计算机视觉方面也获得很大成功 [30-35]，并已经成为一种处理视觉分类问题的基准方法。同时，在视觉方面的研究也促进了多核学习的发展。例如，杨晶晶（音译，Jingjing Yang ）[30]发现全局核权重学习方式在面对视觉数据复杂的分布形态时不能很好地适应，而样本敏感的多核学习[34]又会对噪声产生过度的响应从而导致过拟合和模型退化，并针对这一问题提出组敏感的核权重学习思想。事实上，多核学习的本质，仍旧是特征选择和多源信息融合。和传统的相关方法的研究不同，以多核学习方式进行的信息融合涉及到众多特征之间的结构化信息。今后，这方面的研究仍将是热点。

2.2 跨媒体数据关联建模

在过去的十几年研究当中，为了有效组织网络数据使用户能够准确和快速地检索到具有视觉和语义相关性的网络文档，相关领域的学者从不同几个方面进行了大量的研究工作，例如索引[4]、检索模型[5-10]。最典型的一种适合于大规模数据检索的技术是近似近邻查找技术。例如：局部敏感哈希方法(LSH)[5]被提出以解决高维空间中的近似近邻查找问题。为了进一步提升性能，学者们进一步研究基于学习的哈希方法，例如谱哈希[9]、语义哈希[10]和针对特定任务的哈希码学习技术[11]等。为了利用数据的非线性相似性度量，库里斯（Kulis）等人[12]提出在给定的核表示上直接构建哈希函数，这种技术被称作核化哈希。王树徽等人[13]将核化哈希扩展到多特征表示上。刘威（音译，Wei Liu）等人提出利用样本类信息的基于学习的核化哈希方法[14]。其他一些工作[7][10]提出了一系列算法框架，利用样本类信息和多特征表示进行哈希函数学习。这些工作都是仅仅考虑了单模态数据，并不适用于解决跨模态数据的问题。

本质而言，跨模态数据检索需将不同模态的异构数据映射到一个统一的可度量的表示空间当中。为达到这个目的，两个要求十分重要：首先，在模态内部，语义上相似(不相似)的数据在统一表示空间中也应该相似(不相似)，这种模态内部的相似性可以由局部邻接结构[16][17]或者样本类信息提供[16]。第二，跨模态的相关(不相关)内容在统一表示空间中应该相似(不相似)[17][24]。为达到这两个要求，相关的研究可粗略划分为子空间学习和话题模型两大类。

子空间学习的目的是找到两个模态中使其模态间相关性最大的低维子投影空间表示。经典相关分析(CCA)[18]及其变种[19]提供了一种对这个问题的直接解决方案。 拉斯瓦夏（Rasiwasia）等人基于 CCA 子空间表示提出一种跨模态内容的话题分类器[20]如图 1 所示。其基本流程如下：首先，基于图像和文本文件的共生关系，通过 CCA 学习生成一对使图像和文本内容相关性最大化的子空间，并将图像和文本投射到子空间当中；其次，在各自的子空间表示上构建语义分类器，得到不同模态文件在一个低维语义空间上的概率化表示，这个表示被认为能够很好地体现数据的语义信息；最后，在语义空间上对比不同的跨模态数据之间的语义相关性。然而，该方法忽略了模态内部数据之间的相关性，并且其采用的分步式映射学习策略不能保证所得到的语义映射是最优的，故只能处理小规模跨模态数据。

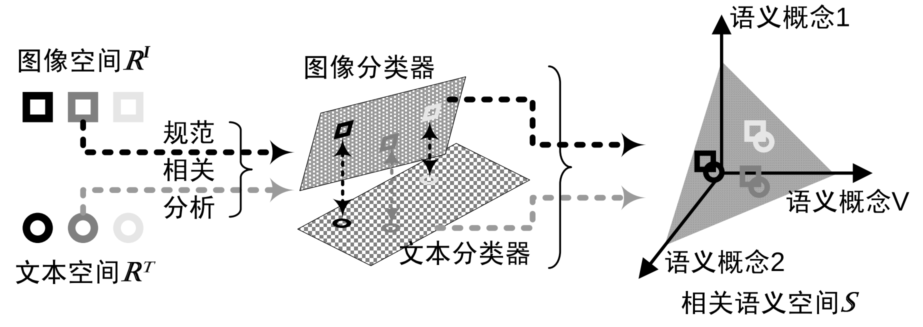


图1. 跨模态关联学习示例

此外，布朗斯廷（Bronstein）等人提出一种基于 boosting 的哈希码学习方法，学习到一系列的“弱哈希函数”及其组合权重，并用来计算跨模态的加权汉明距离[17]。马希（Masci）等人[16]扩展了引文[17]中的模型，在多层神经网络的基础上对模态内部的相似性信息和模态间的相关信息加以利用。基于图表示的方法将模态内部的相似性信息和模态间的相关信息用统一的图结构来表示，而该图表示的最小特征值对应的特征空间就是需要寻找的跨模态子空间。基于类似的思路，宋静宽（音译，Jingkuan Song）等人[44]提出一种基于模态内和模态间关系建模的图分解方法用于跨模态哈希学习。子空间学习的方法一般需要多模态数据严格对齐，同时被组织成一对一的数据对，也就是说，每个文本/视觉文件必须有一个对应的视觉/文本文件。然而，当处理网络数据时，这种要求一般很难满足。另外，子空间学习一般只能针对两个模态的数据，对于多个模态，一般将其分解为一系列的两两模态对应问题，不可避免地带来计算复杂度的提高。

在隐含话题模型中，需要学习隐含话题来对多模态内容的关联方式进行概率化建模。Correspondence LDA (Corr-LDA，一致性隐形狄利克雷分布)方法[21]试图捕捉图像和文字标注之间的话题级别的关系。萧寒（音译，Han Xiao）等人[22]结合 LDA 和 Corr-LDA 等方法用于将图像和声音通过文本标签关联起来。贾扬清（Yangqing Jia）等人提出的模型[23]可以看作是在 LDA 话题模型基础上构建的马尔可夫随机场，其特点是不需要数据以一对一的方式加以组织。甄宜（音译，Yi Zhen）等人[24]提出一种隐含二值嵌入的方法，其本质是同时学习隐含话题分布及二值化的权重表示，并以此来刻画被观测到的模态内部和不同模态数据的相似性。陈宁（音译，Ning Chen）等人[25]提出一种多视角最大间隔（margin）的隐含子空

间学习，获得了非常好的学习效果。然而，虽然有些研究试图在对多模态数据中复杂的话题级别的关系进行良好的建模，这类方法一般不适用于大数据学习问题。

2.3 跨媒体网络社群分析

网络社群的出现是跨媒体兴起的主要原因之一，这为跨媒体数据分析提供了大量信息，为个性化服务提供新契机，但也对计算机领域提出了新的难题。对推特（Twitter）、脸谱（Facebook）等社交网络的分析与研究已经吸引了大量学者。 在推特平台上，研究者从用户交互结果、信息内容和信息时效性等不同角度进行了统计来分析用户行为[26]。一些相关工作进一步展开，例如杨磊等人[27]对 Hashtag（哈希标签）信息传播进行分析建模, 戈什（Ghosh）[28]等通过对推特的链接耕作模式(link farming)进行发掘，从不同角度切入来分析用户的行为。庄金锋（音译，Jingfeng Zhuang）[29]等提出一种面向网络社群用户的融合视觉、文本、社会标记、用户偏好等信息的跨媒体推荐方法。然而，针对网络社群的研究工作仅仅是刚起步，尤其是网络社群和跨媒体内容之间的交互影响机制还不能有效进行分析，还需要学者们更加深入地挖掘与探讨。

3 本课题组的研究工作进展

本组围绕跨媒体数据关联理解与深度挖掘这个科学问题展开了研究工作，内容包括针对跨媒体数据呈现的多态性、异构性、海量性和社会性等特点，分析跨媒体数据中蕴含的热点话题及重大事件结构模式；研究跨媒体数据的语义关联学习方法，建立跨媒体事件的检测、表示和追踪模型；提出检测突发性热点话题及重大事件的计算模型和学习方法，形成基于群体智能的协同反馈计算手段。我们以现实环境的跨媒体数据形态为研究背景，按照“跨媒体语义分析和理解”、“跨媒体话题和事件分析”以及“跨媒体社会属性”等三条主线展开研究工作并推动其不断深入，重点研究如何构建有效的跨媒体语义单元学习模型和不同模态之间的数据关联机制，并利用多模态融合及多源信息（超链接信息、标注信息、社会标签、网络指导信息以及社会群体信息等）提高对跨媒体事件和话题的分析效果。

在跨媒体语义学习方面，我们提出了半监督多核学习方法，以有效应对跨媒体的多样性特点和噪声，在效率和可扩展性上优于现有半监督学习方法；提出了字典学习和判别学习模型，对视觉信息的空间上下文进行建模，有效挖掘层次化语义信息，构建层次化语义标注模型，创新性地提出了一种多层判别字典学习和判别学习交互提升的学习框架；提出了视觉语义关联方法，有效克服了视觉多义性和语义多态性等问题，建立了符合现实跨媒体数据特性的数据库和评测平台，为跨媒体关联分析提供了新的解决思路。

在热点话题和重大事件检测方面，我们提出了跨媒体相似性度量学习方法，以应对跨媒体信息的多模态性，实现了多种不同跨媒体学习任务的信息共享，从而提升了跨媒体相似性度量的表示能力；提出了基于多源信息融合的话题检测模型，对跨媒体事件和话题的社会信息、指导信息、时序信息和多模态信息等进行了有效建模，克服了传统的基于单源信息的话题检测方法的不足。

在跨媒体社会属性分析方面，我们针对移动用户和社会网络用户，提出了若干基于多源属性行为建模的多解析度和结构化行为数据分析、社群发现和实体链接方法，有效应对了群体用户行为的复杂性和多样性。

3.1 跨媒体语义分析和理解

由于海量跨媒体数据的复杂内在分布，跨媒体语义单元学习面临着跨媒体数据特征和高

层语义缺乏一致性、标注数据匮乏、对噪声不够鲁棒、模型可扩展性差、以及跨平台和跨模态的数据分布复杂、关联多样等主要挑战。这些挑战一方面导致现有的特征表示和判别模型不能够很好地适应不同跨媒体语义学习任务的要求，另一方面使得现有跨媒体特征表示不能有效应对跨媒体数据的模态异构性，不利于挖掘其复杂关联关系。针对这些问题，课题组分别在特征表示、判别模型、检索模型上提出一系列行之有效的解决方案，研究成果发表在《美国电机电子工程师学会图像处理汇刊（IEEE Transactions on Image Processing）》、《美国电机电子工程师学会多媒体汇刊（IEEE Transactions on Multimedia）》、美国电机电子工程师学会计算机视觉与模式识别会议（IEEE Conference on Computer Vision and Pattern Recognition，CVPR）等高水平国际期刊和国际会议上。

3.1.1 特征表示

主流的图像描述是基于尺度不变特征转换(Scale-invariant feature transform, SIFT)的视觉词袋模型，但由于其缺乏空间信息描述能力，而且不够紧致，不能很好满足目前对特征的强描述能力和快速高效计算的要求。本课题提出了一种基于结构纹理和边缘提取的紧凑编码模式 Edge-SIFT。为了使生成的 Edge-SIFT 更加紧致，我们提出了二值化压缩和基于 Rankboost 的判别学习方法，以便对该紧凑模式进行选择，得到适应海量近似图像检索任务的紧凑码本。此外，本课题基于所提出的 Edge-SIFT 发展了一种可快速在线验证的倒排索引框架，通过大量实验验证了其有效性和高效性。

在图像语义理解中，视觉多义性和语义多态性问题一直都是一个挑战。视觉多义性是指一块视觉表观可能有很多不同的语义解释；语义多态性是指一个概念在不同的实例下可能有各种不同的视觉表观。本课题提出通过一种新的视角—Vicept 来理解图像，每一个 Vicept 单词是关于一个视觉表观的多概念概率估计。在 Vicept 词典中，每个视觉表观和每个确定的概念都有一个概率联系，这种联系整合在一起可以构成一个视觉表观隶属度的概率分布。为了通过学习，生成有判别能力且结构稀疏的 Vicept，在视觉表观的概念隶属度分布的学习中采用了混合范式正则方法。此外，针对 Vicept 的多层次结构，本课题引入了一种新的距离度量方法，即通过多层次的独立性分析来融合不同层次的 Vicept 描述。

3.1.2 判别模型

针对传统半监督学习方法的不足，我们提出一种可扩展的半监督多核学习方法(S^3MKL)。其损失函数当中包含了有标注训练样本上的训练损失、组稀疏参数正则化和无标注样本上的（组）条件期望一致性损失。与传统的直推式方法不同，所得到的判别模型具有较强的判别性，能够有效预测未知样本的类别标签。在利用海量跨媒体数据进行学习时，数据中蕴含的噪声样本会对判别模型的判别性造成一定的干扰。为了对海量无标注样本进行样本选择，我们基于核化局部敏感哈希方法构建了一个多核哈希系统(MKLSH)，对局部敏感哈希（KLSH）进行了改进，即将在多个核上进行的 KLSH 的汉明码拼接到一起，形成了对海量图像的多核局部敏感哈希表示。在我们的工作中，将总体的半监督多核学习与基于多特征的核化哈希样本选择结合了起来。实验表明这种方法能够更加有效地利用海量无标注样本进行半监督模型学习，并在多个基准数据库上获得了比传统半监督学习方法更佳的分类性能。

传统的图像分类算法往往针对较少类别。但是，现实世界跨媒体数据的类别极多。本课题提出了一种基于树结构的多层判别字典学习算法(ML-DDL)，用于克服现有码本（特征）学习不能有效应对海量类别分类的问题。我们以根据标签信息的语义相关性构建的语义树结构作为先验，通过训练得到一组有监督的码本和分类器模型，利用层次结构进行码本学习，

将原始的极多类问题分解为多个较易处理的多层分类子问题来逐一求解，大大降低运算复杂度，使得有监督的码本学习能适用于海量类别的分类任务，在可承受的时间开销下得到较好的分类性能。

3.1.3 检索模型

为了克服近邻方法的不足，我们提出一种新的近邻相似性度量方法，与以往距离度量的不同之处在于它同时利用了数据的局部密度信息和语义信息。其次，采用基于核化局部敏感哈希方法的多特征近邻搜索策略。最后，为了提高对海量内容的鲁棒性，采用了多特征融合的方法，将在不同特征通道上计算的近邻相似性度量进行融合。在三个经典大规模图像数据库上的大量实验表明，这个方法比传统的近邻方法在语义分析和检索的性能上有较大提升。

为研究跨模态相关模型和跨模态检索技术，我们设计了一套自动数据收集算法来构建跨模态数据库。数据库包括 75K 段文本文档和 35K 幅图像。数据库中话题内容的分布广泛，不同模态的文件数量不均衡，跨模态共生性信息较稀疏，更接近真实跨模态数据。库中包含网页的超链接信息和人工标注的类别信息(预定义的 11 大类)。对图像文件，提取 9 种常用的视觉特征(约 2 万维)，对文本提取经典的 TF-IDF 特征(约 7 万维)。该数据库可用于经典跨模态分析方法的评测以及新的跨模态分析方法的研究和评测。

3.2 跨媒体事件和话题分析

跨媒体事件和话题检测与分析面临着三大挑战：社会交互多样化，新式样层出不穷；数据模态多变，内在关联稀疏；指导信息不足，粒度大小不一。为了对跨媒体事件和话题进行有效表示、检测及追踪，本课题充分考虑跨媒体数据的产生、扩散和关联机制，从如下思路展开研究。第一，利用多特征互补信息以及最大间隔（maximum margin）学习等策略，学习和构建跨媒体话题的相似性度量。第二，融合多源、多模态信息构建跨媒体数据的关联模型，利用热搜词指导发现社会热点话题。研究成果发表在 2012 年的美国电机电子工程师学会计算机视觉与模式识别会议、美国计算机协会 2012 年多媒体会议（ACM Multimedia 2012）、美国计算机协会国际多媒体会议及博览会（IEEE ICME）等国际会议上。

3.2.1 跨媒体结构表示

对于海量跨媒体信息处理的研究而言，寻求理想的距离度量表示是绝大多数分析模型的核心部分或者研究重点。然而，传统的度量学习方法无法很好地适应高维多特征表达以及复杂的语义结构和表观视觉分布。为此，我们提出了一种有效的多任务多特征度量学习方法，利用网络跨媒体的语义标注信息和社会标签信息进行多任务学习，得到一种在多特征表示下具有语义一致性的低复杂性度量准则。所提方法能够有效融合多种特征表示，相比于传统方法，学习到的特征的复杂度(模型参数个数)也显著降低。该方法的另外一个优点是能够根据学习任务的需求，有效控制需要优化的相似性度量数量，在准确率和训练时间开销之间的折衷可达到更好效果。在多个数据库上的多项实验表明该方法的性能比其它方法有显著的提高。

3.2.2 基于多源信息融合的跨媒体事件和话题分析

不同于传统的基于聚类的主题检测方法，我们提出了一种新颖的基于多线索融合的网络视频话题检测方法。首先，利用与视频相关的标签信息，提取密集突发的标签组，作为事件的备选。其次，检测相似视频片段，并将其与视频的标签进行融合形成视频标签组。最后，通过对热搜词的时域特征分析，过滤掉突发性低的热搜词，指导事件检测。

传统的话题检测方法大多只能处理单一媒体的数据源，其信息量、受众、关注点往往是有限的。相比之下，来自于不同媒体源的信息能够互相补充，信息量更加丰富，能更好地反映社会现实。因此，有效利用不同数据源间的互补性，是提升话题检测与跟踪性能的有效途径。为此，我们提出一种灵活的多模态信息融合的跨媒体数据表示框架,充分利用跨模态数据间的语义关联信息,对跨媒体中的话题结构进行检测。首先，建立多模态图，图中边的权重融合了多模态内容的相似性和时间信息。由于属于同一话题的数据自然地形成具有紧密相似度关系的密集子图，故可通过图聚类方法查找密集子图，从而实现跨媒体话题检测。在公共数据集及自建跨媒体数据集上的实验结果表明这一策略能够有效检测跨媒体话题。

3.3 跨媒体用户行为分析

针对移动用户和社会网络用户，我们提出了若干基于多源属性行为建模的轨迹数据分析和社群发现方法，有效应对了群体用户行为的复杂性和多样性。具体而言，我们针对大规模的用户轨迹行为数据，分别提取轨迹中地点的语义信息、速度模式信息、时间间隔模式信息和轨迹物理相似性信息，最后将多种行为的相似度进行加权融合，并利用密集子图检测方法检测到一系列具有长时段相似行为的用户群落(communities)。针对社会网络用户，引入多媒体内容分析技术，提出一种多源异构行为建模框架，对属性信息（性别、邮箱、国籍、年龄等）提出一种概率化匹配方法；对用户的行为倾向性，提出一种多时域解析度的内容分布描述方法。对用户的转载、引用、地点记录等行为模式，提出一种多时间窗宽的匹配框架，并利用神经网络的池化方法去计算用户在多个时间窗宽度上的总体行为相似性。基于这些行为相似性描述，提出一种基于多目标优化的结构化匹配学习方法，有效利用社会网络中用户的好友信息对判别结果进行有效扩散，达到了对跨社会媒体平台用户进行自动匹配的目的。研究成果分别被美国计算机协会的国际信息与知识管理大会（ACM CIKM 2013）和数据管理专业委员会年度会议（SIGMOD 2014）录用。

3.4 研究进展小结

长期以来，针对单一媒体和多媒体的事件和话题分析，国内外研究机构往往采用单一媒体主题建模、单一模态分析方法和简单的话题结构建模方法。本课题充分考虑了跨媒体数据呈现的多态性、异构性、海量性和社会性特点，以现实环境的跨媒体数据形态为研究背景，按照“跨媒体语义单元学习”、“热点话题和事件检测”和“跨媒体用户行为分析”三条主线开展研究，基于多源信息和多特征融合这个主要的研究出发点，有效利用跨媒体上下文信息，构建适合海量跨媒体数据的语义分析、内容理解和关联框架，解决了在高噪声和复杂关联背景下对现实跨媒体的语义分析、事件话题分析和用户行为分析等问题。

与国内外同类研究工作相比，本实验室的主要创新性成果包括：提出半监督多核学习方法，利用跨媒体数据源的多特征表示，有效应对跨媒体的多样性特点和噪声问题，克服了传统的基于多特征融合的语义学习方法的不足；提出字典学习和判别学习模型，对视觉信息的空间上下文进行建模，有效挖掘层次化语义信息，构建层次化语义标注模型，并针对层次化语义类别结构，创新性地提出了一种多层判别字典学习和判别学习交互提升的学习框架；提出视觉语义关联方法，有效克服了视觉多义性和语义多态性等带来的困难，构建了高维视觉数据到语义空间的映射模型，建立了符合现实跨媒体数据特性的数据库和评测平台，为跨媒体关联分析提供了新的解决思路；提出跨媒体相似性度量学习方法，构建了低复杂度的跨媒体相似性度量，满足了复杂跨媒体学习任务的需求，并实现了多种不同跨媒体学习任务的信息共享，从而提升了跨媒体相似性度量的表示能力；提出基于多源信息融合的话题检测模型，对跨媒体事件和话题的社会信息、指导信息、时序信息和多模态信息等进行了有效建模，克服了传统的基于单源信息的话题检测方法的不足，实现了跨媒体话题检测。在跨媒体社会网

络用户行为方面，提出了有别于以往基于纯文本行为分析的一系列基于多源行为分析的方法，有效地解决了社会网络社群发现和账户链接等应用问题，为社会网络行为分析提供了一种新的研究思路。

4 总结

在未来五到十年内，跨媒体分析和检索技术将逐渐成为学术界和产业界的研究热点。由于跨媒体大数据中蕴含着极大的价值，能否有效地挖掘这些价值，将直接决定各类信息和知识服务系统的服务质量和用户体验满意度，决定媒体大数据分析产业的兴衰成败。

从未来的发展趋势来看，跨媒体分析的核心目标仍然将是“语义”、“关联”和“社群”。为适应跨媒体数据自身各种复杂的特性，在数据分析理论上亟需更具有指导性和针对性的理论方法，相应的分析和检索技术也必须不断创新，才能够更好地满足日益增长的媒体大数据分析的需求。

参考文献：

[1] Smeulders A. W. M., Worring M., Santini S., Gupta A., Jain R.. Content-based image retrieval at the end of the early years. PAMI, 22(12): 1349 – 1380, 2000

[2] Shechtman E. and Irani M.. Matching local self-similarities across images and videos. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1-8 (2007)

[3] Hauagge D.C. and Snavely N.. Image matching using local symmetry features. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 206-213 (2012)

[4] Nister D. and Stewenius H.. Scalable recognition with a vocabulary tree. In CVPR, 2006.

[5] Datar M., Immorlica N., Indyk P., and Mirrokni V. S.. Locality-sensitive hashing scheme based on *p*-stable distributions. In Proceedings of the 20th annual symposium on Computational Geometry, pages 253 – 262, 2004.

[6] Weiss Y., Torralba A., and Fergus R.. Spectral hashing. In NIPS, 2008.

[7] Song J., Yang Y., Huang Z., Shen H., and Hong R.. Multiple Feature Hashing for Real-time Large Scale Near-duplicate Video Retrieval. In ACM Multimedia, 2011

[8] Zhang D., Wang J., Cai D., and Lu J.. Self-taught Hashing for Fast Similarity Search. In SIGIR, 2010.

[9] Weiss Y., A. Torralba, and R. Fergus. Spectral hashing. In NIPS, 2008.

[10] Salakhutdinov R. and Hinton G.. Semantic hashing. International Journal of Approximate Reasoning, 50(7), 2009

[11] Shakhnarovich G.. Learning task-specific similarity, 2005. PhD thesis, MIT.

[12] Kulis B. and Grauman K.. Kernelized locality sensitive hashing for scalable image search. In ICCV, 2009.

[13] Wang S., Huang Q., Jiang S., and Tian Q.. S^3MKL: Scalable semi-supervised multiple kernel learning for real-world image applications. TMM, 14(4):1259–1274, 2012.

[14] Liu W., Wang J., Ji R., Jiang Y.-G., and Chang S.-F.. Supervised hashing with kernels. In CVPR, 2012

[15] Zhang D., Wang F. and Si L.. Composite hashing with multiple information sources. In SIGIR, 2011.

[16] Masci J., Bronstein M. M., Bronstein A. M., and Schmidhuber J.. Multimodal similarity-preserving

hashing, 2012. arXiv:1207.1522

[17] Bronstein M. M., Bronstein A. M., Michel F., and Paragios N.. Data fusion through cross-modality metric learning using similarity-sensitive hashing. In CVPR, 2010.

[18] Hotelling H.. Relations between two sets of variates. Biometrika, 28(34):321–372, 1936.

[19] Chen X., Liu H., and Carbonell J. G.. Structured sparse canonical correlation analysis. In AISTATS, 2012.

[20] Rasiwasia N., Pereira J., Coviello E., Doyle G., Lanckriet G., Levy R., and Vasconcelos N.. A new approach to cross-modal multimedia retrieval. In ACM Multimedia, 2010

[21] Blei D. and Jordan M.. Modeling annotated data. In SIGIR, 2003.

[22] Xiao H. and Stibor T.. Toward artificial synesthesia: Linking images and sounds via words. In NIPS workshop on Machine Learning for next generation Computer Vision challenges, 2010.

[23] Jia Y., Salzmann M., and Darrell T.. Learning cross-modality similarity for multinominal data. In ICCV, 2011.

[24] Zhen Y. and Yeung D.-Y.. A probabilistic model for multimodal hash function learning. In KDD, 2012.

[25] Chen N., Zhu J., Sun F., and Xing E. P.. Large-margin predictive latent subspace learning for multi-view data analysis. TPAMI, 34(12):2365–2378, 2012.

[26] Kwak H., Lee C., Park H., and Moon S.. What is Twitter, a Social Network or a News Media? In WWW, 2010.

[27] Yang L., Sun T., and Mei Q.. We Know What @You #Tag: Does the Dual Role Affect Hashtag Adoption? In WWW, 2012.

[28] Ghosh S., Viswanath B., Kooti F., Sharma N., Gautam K., Benevenuto F., Ganguly N., and Gummadi K.. Understanding and Combating Link Farming in the Twitter Social Network. In WWW, 2010.

[29] Zhuang J., Mei T., Hoi S. C. H., Hua X.-S., Li S.. Modeling social strength in social media community via kernel-based learning. ACM Multimedia 2011: 113-122.

[30] Yang J., Li Y., Tian Y., Duan L., and Gao W.. Group Sensitive Multiple Kernel Learning for Object Categorization. ICCV, 2009.

[31] Varma M. and Ray D.. Learning the Discriminative Power-invariance Trade-off. ICCV, 2007.

[32] Vedaldi A., Gulshan V., Varma M., and Zisserman A.. Multiple Kernels for Object Detection. ICCV, 2009.

[33] Bucak S., Jin R. and Jain A.. Multi-label Multiple Kernel Learning by Stochastic Approximation: Application to Visual Object Recognition. NIPS, 2010.

[34] Cao L., Luo J., Liang F., and Huang T.. Heterogeneous Feature Machine for Visual Recognition. *ICCV*, 2009.

[35] Liu J., Ali S. and Shah M.. Recognizing human actions using multiple features. *CVPR*, 2008.

[36] Yang, J., Yu, K., Gong, Y., & Huang, T. Linear spatial pyramid matching using sparse coding for image classification. In Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on (pp. 1794-1801).

[37] Wang, J., Yang, J., Yu, K., Lv, F., Huang, T., & Gong, Y. Locality-constrained linear coding for image classification. In Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on (pp. 3360-3367).

[38] Mairal J, Bach F, Ponce J. Task-driven dictionary learning. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 2012, 34(4): 791-804.

[39] Deng J, Krause J, Berg A C, et al. Hedging your bets: Optimizing accuracy-specificity trade-offs in large scale visual recognition. Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on. IEEE, 2012: 3450-3457.

[40] Grauman K, Sha F, Hwang S J. Learning a tree of metrics with disjoint visual features. Advances in Neural Information Processing Systems. 2011: 621-629.

[41] Hinton G E, Salakhutdinov R R. Reducing the dimensionality of data with neural networks. Science, 2006, 313(5786): 504-507.

[42] Krizhevsky A, Sutskever I, Hinton G. Imagenet classification with deep convolutional neural networks. Advances in Neural Information Processing Systems. 2012: 1106-1114.

[43] Xiao L, Zhou D, Wu M. Hierarchical classification via orthogonal transfer. Proceedings of the 28th International Conference on Machine Learning (ICML-11). 2011: 801-808.

[44] Song J, Yang Y, Yang Y, et al. Inter-media hashing for large-scale retrieval from heterogeneous data sources. Proceedings of the 2013 international conference on Management of data. ACM, 2013: 785-796.

作者简介：

王树徽： 中国科学院计算技术研究所，智能信息处理重点实验室，博士后 wangshuhui@ict.ac.cn

黄庆明： 中国科学院计算技术研究所，智能信息处理重点实验室，客座教授，博士生导师

（ 上接第34页 ）

[78] Mansour Y, Mohri M, Rostamizadeh A. (2009). Domain Adaptation: Learning Bounds and Algorithms. *Proceedings of 22nd Annual Conference on Learning Theory*, San Francisco: Morgan Kaufmann.

[79] Zhuang, F Z, Luo, P, Shen, Z, et al. (2010) Collaborative dual-plsa: mining distinction and commonality across multiple domains for text classification. *Proceedings of the 19th ACM international conference on Information and knowledge management*. ACM:359-368.

[80] Xue G R, Dai W Y, Yang, Q, et al. (2008). Topic-bridged PLSA for Cross-domain Text Classification. *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, New York, NY, USA: ACM: 627-634.

[81] Wang H, Huang H, Nie F, et al. (2011). Cross-language web page classification via dual knowledge transfer using nonnegative matrix tri-factorization. *Proc. of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*. ACM: 933-942.

[82] Zhuang F Z, Luo P, Yin P F, et al. (2013). Concept learning for cross-domain text classification: a general probabilistic framework. *Proceedings of the Twenty-Third international joint conference on Artificial Intelligence*, AAAI Press:1960-1966.

作者简介：

庄福振 中国科学院计算技术研究所智能信息处理重点实验室 副研究员，zhuangfz@ics.ict.ac.cn

何 清 中国科学院计算技术研究所智能信息处理重点实验室 研究员